Procesamiento del Lenguaje Natural, núm. 30 (2003), pp. 115-116 recibido 1-03-2003; aceptado 20-05-2003

Automatización del análisis sintáctico del español

Luis Javier Losada García
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
losada@dis.ulpgc.es

Director: Octavio Santana Suárez
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
osantana@dis.ulpgc.es
http://www.gedlc.ulpgc.es

Codirector: José Rafael Pérez Aguiar
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
jperez@dis.ulpgc.es

Resumen: En esta tesis se desarrolla una solución para la automatización del análisis sintáctico del español, se identifican los principales problemas que aparecen en la automatización, se proponen soluciones a los mismos, se aportan métodos para la desambiguación funcional y estructural. Para demostrar su funcionalidad se han desarrollado dos aplicaciones: un desambiguador funcional y un analizador morfosintáctico del español.
Palabras clave: desambiguación, sintaxis, análisis sintáctico, lingüística computacional.

Abstract: The objetive of this thesis is to present some techniques, which deals with the automatic syntactic analysis of Spanish texts. In the present work we deal with problems of automatic processing, we suggest some solutions, and we develop some programs in order to show how to realice the functional and structural disambiguation. We develop two programs that apply the ideas and programming paradigms explained in this thesis: a functional disambiguation program and a morphosyntactic analysis program.
Keywords: disambiguation, syntax, syntactic analysis, computational linguistic.

La presente tesis prolonga la línea de los trabajos realizados por el Grupo de Estructuras de Datos y Lingüística Computacional de la ULPGC durante los últimos años, en el ámbito de las aplicaciones orientadas a la lingüística computacional y al procesamiento del lenguaje natural. A partir de los trabajos de reconocimiento y generación morfológica automáticos se pasa al siguiente nivel en el campo de la lingüística: la sintaxis. En esta tesis se han logrado soluciones a los problemas que se producen a la hora de realizar la automatización de la sintaxis.

Se ha realizado un amplio estudio de la gramática española que dio lugar a un conjunto finito de reglas de representación. Los problemas de representación debidos al enorme número de combinaciones posibles han sido resueltos mediante la aplicación de reglas especiales. Posteriormente han sido desarrollados mecanismos de eliminación de ambigüedades que se producen durante los procesos de generación de árboles de representación sintáctica y se han puesto a punto técnicas para la optimización de los algoritmos de generación de árboles.

Las posibles ambigüedades se desglosan en dos grupos: funcionales —aquellas debidas a la multiplicidad de respuestas del analizador morfológico para una misma voz— y las estructurales —debidas a la multiplicidad de árboles de representación para una misma combinación de categorías funcionales. La mayoría de las ambigüedades funcionales se resuelven de manera local, mediante algoritmos que se basan en qué elementos se combinan en los diferentes sintagmas posibles y bajo qué condiciones lo hacen —desambiguador funcional local.

ISSN 1135-5948

La generación de árboles de representación sintáctica consiste en construir árboles a partir de las categorías funcionales de las palabras de una sentencia —cualquier árbol o subárbol que se construye cubre una combinación de categorías funcionales entre dos posiciones de la sentencia. Ya que en la generación de los árboles de representación sintáctica sólo se contemplan las combinaciones de categorías funcionales devueltas por el desambiguador funcional local, se consigue eliminar un amplio conjunto de combinaciones de partida que daría lugar a una enorme cantidad de subárboles de representación no correctos.

Se aplica un conjunto de reglas de desambiguación estructural que realizan podas durante la generación de los árboles de representación sintáctica. Estas reglas se basan en las relaciones entre los diferentes elementos de la oración, tanto si se trata de relaciones entre dos estructuras sintácticas separadas o de una estructura sintáctica con los elementos que han dado lugar a la misma —palabras, categorías funcionales y otros elementos sintácticos—; también se ha introducido cierta información semántica a partir de información propia de diccionarios ideológicos y de ideas afines.

Como resultado de los trabajos realizados se han desarrollado dos motores *DeFuSE* —Desambiguador Funcional de Sentencias del Español— y *AMoSinE* —Analizador MorfoSintáctico del Español. En ambos casos se han desarrollado interfaces de usuario que dan lugar a dos aplicaciones finales.

DeFuSe se orienta a los procesos de desambiguación funcional local y sirve como base para las aplicaciones de niveles superiores. La desambiguación funcional local se ha desarrollado en función de un estudio de las relaciones de vecindad en el seno de las estructuras básicas de la gramática española: los sintagmas. Permite seleccionar qué procesos se realizan durante la desambiguación a fin de observar la influencia de los distintos aspectos considerados; de esta forma puede tener en cuenta o no: las relaciones de vecindad, la concordancia en la flexión, las combinaciones que serían admitidas por encontrarse en los límites de las estructuras locales y que no deben aceptarse —combinaciones vedadas— y los casos especiales que se derivan del estudio de grupos concretos de palabras. Por otro lado, admite una serie de restricciones que descartan opciones infrecuentes como las interpretaciones antiguas o en desuso de las palabras, los otros valores que puedan tener los artículos, las conjunciones, las preposiciones y los pronombres y considerar sólo las combinaciones funcionales que contengan formas verbales.

AMoSinE se trata de un motor orientado al análisis de las estructuras sintácticas de una sentencia, con capacidad para ponerlas a disposición de otras aplicaciones. Se basa en una definición en forma de reglas simples de la gramática española; se utilizan durante el proceso de generación de árboles de representación mediante un algoritmo de tipo 'chart' al que se le han añadido una serie de mecanismos de poda. La entrada está constituida por los resultados de *DeFuSe* y no por la sentencia en sí. En *AmoSinE* se define un conjunto de opciones que pueden ser activadas o desactivadas para ver los efectos de los grupos de reglas sobre la desambiguación estructural; así, es posible realizar el análisis para todas las combinaciones funcionales o sólo para las devueltas por *DeFuSe*, aplicar información semántica o puramente estructural. También se pueden activar o desactivar las diferentes restricciones expuestas en la desambiguación funcional local respecto de: las oraciones verbales, las interpretaciones antiguas o desusadas, los artículos, las conjunciones, las preposiciones y los pronombres.

Uno de los principales problemas a la hora de implementar sistemas automáticos de análisis sintáctico radica en el enorme costo computacional que conlleva el tratamiento de la información necesaria. Para paliar este problema se introduce el concepto de pre-regla: consiste en adelantar lo más posible la acción de las reglas de desambiguación estructural, ya que se pueden rechazar símbolos aún sin conocer todos los que lo producirían. De este modo, se introduce un sistema de poda predictiva que ha mostrado reducir los tiempos de análisis.

En cuanto a los resultados estadísticos obtenidos sobre texto real, el número de combinaciones funcionales se reduce como promedio tras el proceso de desambiguación funcional local en alrededor del 86% y tras la desambiguación estructural, en torno al 95%. Se concluye que para llegar a un análisis completo que no presente ambigüedades en la respuesta será necesario introducir una mayor información semántica.